One-Dimensional Partitioning for Heterogeneous Systems: Theory and Practice

Ali Pınar [a,1], E. Kartal Tabak [b] and Cevdet Aykanat [b,2]

[a] *Computational Research Division. Lawrence Berkeley National Laboratory*

[b] *Department of Computer Engineering, Bilkent University*

Abstract

We study the problem of one-dimensional partitioning of nonuniform workload arrays with optimal load balancing for heterogeneous systems. We look at two cases: chain-on-chain partitioning, where the order of the processors is specified, and chain partitioning, where processor permutation is allowed. We present polynomial time algorithms to solve the chain-on-chain partitioning problem optimally, while we prove that the chain partition is NP-complete. Our empirical studies show that our proposed exact algorithms produce substantially better results than heuristics while the solution times remain comparable.

Key words: parallel computing; one-dimensional partitioning; load balancing; chain-on-chain partitioning; dynamic programming; parametric search;

Email addresses: apinar@lbl.gov (Ali Pınar), tabak@cs.bilkent.edu.tr (E. Kartal Tabak), aykanat@cs.bilkent.edu.tr (Cevdet Aykanat).

[1] Supported by the Director, Office of Science, Division of Mathematical, Information, and Computational Sciences of U.S. Department of Energy under contract DE-AC03-76SF00098. One Cyclotron Road MS 50F, Berkeley, CA 94720

[2] Corresponding author. Partially supported by Scientific and Research Council of

1 Introduction

In many applications of parallel computing, load balancing is achieved by mapping a possibly multi-dimensional computational domain down to a one-dimensional (1D) array, and then partitioning this array into parts with equal weights. Space filling curves are commonly used to map the higher dimensional domain to a 1D workload array to preserve locality and minimize communication overhead after partitioning [5,8,13]. Similarly, processors can be mapped to a 1D array so that communication is relatively faster between close processors in this processor chain [9]. This eases mapping for computational domains and improves efficiency of applications. The load balancing problem for these applications can be modeled as the chain-on-chain partitioning (CCP) problem, where we map a chain of tasks onto a chain of processors. Formally, the objective of the CCP problem is to find a sequence of $P-1$ separators to divide a chain of N tasks with associated computational weights into P consecutive parts to minimize maximum load among processors.

In our earlier work [15], we studied the CCP problem for homogenous systems, where all processors have identical computational powers. We have surveyed the rich literature on this problem, proposed novel methods as well as improvements on existing methods, and studied how these algorithms can be implemented efficiently to be effective in practice. In this work, we investigate how these techniques can be generalized for heterogeneous systems, where processors have varying computational powers. Two distinct problems arise in partitioning chains for heterogeneous systems. The first problem is the CCP problem, where a chain of tasks is to be mapped onto a chain of processors,

Turkey (TÜBİTAK) under grant 103E028

i.e., the pth task subchain in a partition is assigned to the pth processor. The second problem is the chain partitioning (CP) problem, where a chain of tasks is to be mapped onto a *set*, as opposed to a chain, of processors, i.e., processors can be permuted for subchain assignments. For brevity, the CCP problem for homogenous systems and heterogeneous systems will be referred to as the homogenous CCP problem and heterogeneous CCP problem, respectively. The CP problem refers to the chain partitioning problem for heterogeneous systems, since it has no counterpart for homogenous systems.

In this article, we show that the heterogeneous CCP problem can be solved in polynomial time by enhancing the exact algorithms proposed for the solution of the homogenous CCP problem [15]. We present how these exact algorithms for homogenous systems can be enhanced for heterogeneous systems and implemented efficiently for runtime performance. We also present how the heuristics widely used for the solution of homogenous CCP problem can be adapted for heterogeneous systems. We present the implementation details and pseudocodes for the exact algorithms and heuristics for clarity and reproducibility. Our experiments with workload arrays coming from image-space-parallel volume rendering and row-parallel sparse matrix vector multiplication applications show that our proposed exact algorithms produce substantially better results than the heuristics while the solution times remain comparable. On average, optimal solutions provide 10.3 and 9.0 times better load imbalance than heuristics for 256-way partitionings of volume rendering and sparse matrix datasets, respectively. On average, the time it takes to compute an optimal solution is less than 2.70 times the time it takes to compute an approximation using heuristics for 256 processors, and thus the preprocessing times can be easily compensated by the improved efficiency of the subsequent

computation even for a few iterations.

The CP problem on the other hand, is NP-complete as we prove in this paper. Our proof uses a pseudo-polynomial reduction from the 3-Partition problem, which is known to be NP-complete in the strong sense [6]. Our empirical studies showed that processor ordering has a very limited effect on the solution quality, and an optimal CCP solution on a random processing ordering serves as an effective CP heuristic.

The remainder of this paper is organized as follows. Table 1 summarizes important symbols used throughout the paper. Section 2 introduces the heterogeneous CCP problem. In Section 3, we summarize the solution methods for homogenous CCP. In Section 4, we discuss how solution methods for homogenous systems can be enhanced to solve the heterogeneous CCP problem. In Section 5, we discuss the CP problem, prove that it is NP-Complete. We present the results of our empirical studies with the proposed methods in Section 6, and finally, we conclude with Section 7.

2 Chain-on-chain (CCP) Problem for Heterogeneous Systems

In the heterogeneous CCP problem, a computational problem, which is decomposed into a chain $\mathcal{T} = \langle t_1, t_2, \ldots, t_N \rangle$ of N tasks with associated *positive* computational weights $\mathcal{W} = \langle w_1, w_2, \ldots, w_N \rangle$ is to be mapped onto a processor chain $\mathcal{P} = \langle \mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_P \rangle$ of P processors with associated execution speeds $\mathcal{E} = \langle e_1, e_2, \ldots, e_P \rangle$. The execution time of task t_i on processor $\mathcal{P}_p$ is w_i/e_p. For clarity, we note that there are no precedence constraints among the tasks in the chain.

A *task subchain* $\mathcal{T}_{i,j} = \langle t_i, t_{i+1}, \ldots, t_j \rangle$ is defined as a subset of contiguous tasks.

Table 1
The summary of important abbreviations and symbols

Notation	Explanation
N	number of tasks
$\mathcal{T}$	task chain, i.e., $\mathcal{T} = \langle t_1, t_2, \ldots, t_N \rangle$
t_i	ith task in the task chain
$\mathcal{T}_{i,j}$	task subchain of tasks from t_i upto t_j, i.e., $\mathcal{T}_{i,j} = \langle t_i, t_{i+1}, \ldots, t_j \rangle$
w_i	computational load of task t_i
$w_{\max}$	maximum computational load among all tasks
w_{avg}	average computational load of all tasks
$w_{\min}$	minimum computational load of all tasks
$W_{i,j}$	total computational load of task subchain $\mathcal{T}_{i,j}$
W_{tot}	total computational load, i.e., $W_{\text{tot}} = W_{1,N}$
P	number of processors
$\mathcal{P}$	processor chain, i.e., $\mathcal{P} = \langle \mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_P \rangle$ in the CCP problem
	processor set, i.e., $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_P\}$ in the CP problem
$\mathcal{P}_p$	pth processor in the processor chain
$\mathcal{P}_{q,r}$	processor subchain from $\mathcal{P}_q$ upto $\mathcal{P}_r$, i.e., $\mathcal{P}_{q,r} = \langle \mathcal{P}_q, \mathcal{P}_{q+1}, \ldots, \mathcal{P}_r \rangle$
e_p	execution speed of processor $\mathcal{P}_p$
$E_{q,r}$	total execution speed of processor subchain $\mathcal{P}_{q,r}$
E_{tot}	total execution speed of all processors, i.e., $E_{\text{tot}} = E_{1,P}$
B^*	ideal bottleneck value, achieved when all processors have load in proportion to their speed
UB	upper bound on the value of an optimal solution
LB	lower bound on the value of an optimal solution
s_p	index of the last task assigned to the pth processor.
$\lg x$	base-2 logarithm of x, i.e., $\lg x = \log_2 x$.

The computational weight of subchain $\mathcal{T}_{i,j}$ is $W_{i,j} = \sum_{h=i}^{j} w_h$. A partition Π should map contiguous task subchains to contiguous processors. Hence, a P-way partition of a task chain with N tasks onto a processor chain with P processors is described by a sequence $\Pi = \langle s_0, s_1, \ldots, s_P \rangle$ of $P+1$ separator indices, where $s_0 = 0 \leq s_1 \leq \cdots \leq s_P = N$. Here, s_p denotes the index of the last task of the pth part so that processor $\mathcal{P}_p$ receives the task subchain $\mathcal{T}_{s_{p-1}+1,s_p}$ with load $W_{s_{p-1}+1,s_p}/e_p$. The cost $C(\Pi)$ of a partition Π is determined by the maximum processor load among all processors, i.e.,

$$C(\Pi) = \max_{1 \leq p \leq P} \left\{ \frac{W_{s_{p-1}+1,s_p}}{e_p} \right\} \tag{1}$$

This $C(\Pi)$ value of a partition is called its *bottleneck value*, and the processor defining it is called the *bottleneck processor*. The CCP problem is to find a partition Π_{opt} that minimizes the bottleneck value $C(\Pi_{\text{opt}})$.

Similar to the task subchain definition, a processor subchain $\mathcal{P}_{q,r} = \langle \mathcal{P}_q, \mathcal{P}_{q+1}, \ldots, \mathcal{P}_r \rangle$ is defined as a subset of contiguous processors. The computational speed of $\mathcal{P}_{q,r}$ is $E_{q,r} = \sum_{p=q}^{r} e_p$.

The ideal bottleneck value B^* is achieved when all processors are equally loaded as

$$B^* = \frac{W_{\text{tot}}}{E_{\text{tot}}}, \tag{2}$$

where E_{tot} is the sum of all processor speeds and W_{tot} is the total task weight; i.e., $E_{\text{tot}} = E_{1,P}$ and $W_{\text{tot}} = W_{1,N}$.

3 CCP Algorithms for Homogenous Systems

The homogenous CCP problem can be considered as a special case of the heterogeneous CCP problem, where the processors are assumed to have equal speed, i.e., $e_p = 1$ for all p. Here, we review the CCP algorithms for homogenous systems. A comprehensive review and presentation of homogenous CCP algorithms is available in [15].

3.1 Heuristics

Possibly the most commonly used CCP heuristic is *recursive bisection* (RB), a greedy algorithm. RB achieves P-way partitioning through $\lg P$ levels of bisection steps. At each level, the workload array is divided evenly into two. RB finds the optimal bisection at each level, but the sequence of optimal bisections at each level may lead to a multi-way partition which is far away from an optimal. Pınar and Aykanat [15] proved that RB produces partitions with bottleneck values no greater than $B^* + w_{\max}(P-1)/P$.

Miguet and Pierson [11] proposed another heuristic that determines s_p by bipartitioning the task chain in proportion to the length of the respective

processor subchains. That is, s_p is selected in such a way that $W_{1,s_p}/W_{1,N}$ is close to the ratio p/P as much as possible. Miguet and Pierson [11] prove that the bottleneck value found by this heuristic has an upper bound of $B^*+w_{\max}$.

These heuristics can be implemented in $O(N+P\lg N)$ time. The $O(N)$ time is due to prefix-sum operation on the tasks array, after which, each separator index can be found by a binary search on the prefix-summed array.

3.2 Dynamic Programming

The overlapping subproblems and the optimal substructure properties of the CCP problem enable dynamic programming solutions. The overlapping subproblems are partitioning the first i tasks onto the first p processors, for all possible i and p values. For the *optimal substructure* property, observe that if the last processor is not the bottleneck processor in an optimal partition, then the partitioning of the remaining tasks onto the first $P-1$ processors must be optimal. Hence, the recursive definition for the bottleneck value of an optimal partition is

$$B_i^p = \min_{0\leq j\leq i} \left\{ \max \left\{ B_j^{p-1}, W_{j+1,i} \right\} \right\} \tag{3}$$

Here, B_i^p denotes the optimal solution value for partitioning the first i tasks onto the first p processors. In Eq. (3), searching for index j corresponds to searching for separator s_{p-1} so that the remaining subchain $\mathcal{T}_{j+1,i}$ is assigned to the last processor in an optimal partition. This definition defines a dynamic programming table of size PN, and computing each entry takes $O(N)$ time, resulting in an $O(N^2P)$-time algorithm. Choi and Narahari [3], and Olstad and Manne [10] reduced the complexity of this scheme to $O(NP)$ and $O((N-P)P)$, respectively. Pınar and Aykanat [15] presented enhancements to limit the search space of each separator by exploiting upper and lower bounds on

the optimal solution value for better practical performance.

3.3 Parametric Search

Parametric search algorithms rely on two components: a probing operation to determine if a solution exists whose bottleneck value is no greater than a specified value, and a method to search the space of candidate values. The probe algorithm can be computed only in $O(P \lg N)$ time by using binary search on the prefix-summed workload array. Below, we summarize algorithms to search the space of bottleneck values.

3.3.1 Nicol's Algorithm

Nicol's algorithm [12] exploits the fact that any candidate B value is equal to the weight of a task subchain. A naive solution is to generate all subchain weights, sort them, and then use binary search to find the minimum value for which a probe succeeds. Nicol's algorithm efficiently searches for this subchain by considering each processor in order as a candidate bottleneck processor. For each processor $\mathcal{P}_p$, the algorithm does a binary search for the smallest index that will make $\mathcal{P}_p$ the bottleneck processor. With the $O(P \lg N)$ cost of each probing, Nicol's algorithm runs in $O(N+(P \lg N)^2)$ time.

Pınar and Aykanat [15] improved Nicol's algorithm by utilizing the following simple facts. If the probe function succeeds (fails) for some B, then probe function will succeed (fail) for any $B' \leq (\geq) B$. Therefore by keeping the smallest B that succeeded and the largest B that failed, unnecessary probing is eliminated, which drastically improves runtime performance [15].

3.3.2 Bidding Algorithm

The bidding algorithm [15,14] starts with a lower bound and proceeds by gradually increasing this bound until a feasible solution value is reached. The increments are chosen to be minimal so that the first feasible bottleneck value is optimal. Consider the partition generated by a failed probe call that loads the first $P-1$ processors maximally not to exceed the specified probe value. To find the next bottleneck value, processors bid with the bottleneck value that would add one more task to their domain, and the minimum bid among the processors is chosen to be the next bottleneck value. The bidding algorithm moves each one of the P separators for $O(N)$ positions in the worst case, where choosing the new bottleneck value takes $O(\lg P)$ time using a priority queue. This makes the complexity of the algorithm $O(NP \lg P)$.

3.3.3 Bisection Algorithms

The bisection algorithm starts with a lower and an upper bound on the solution value and uses binary search in this interval. If the solution value is known to be an integer, then the bisection algorithm finds an optimal solution. Otherwise, it is an ϵ-approximation algorithm, where ϵ is the user defined accuracy for the solution. The bisection algorithm requires $O(\lg(w_{\max}/\epsilon))$ probe calls, with $O(N+P \lg N \lg(w_{\max}/\epsilon))$ overall complexity.

Pınar and Aykanat [15] enhanced the bisection algorithm by updating the lower and upper bounds to realizable bottleneck values (the total weight of a subchain). After a successful probe, the upper bound can be set to be the bottleneck value of the partition generated by the probe function, and after a failed probe, the lower bound can be set to be the smallest value that might succeed, as in the bidding algorithm. These enhancements transform the bi-

section algorithm to an exact algorithm, as opposed to an ϵ-approximation algorithm.

4 Proposed CCP Algorithms for Heterogeneous Systems

The algorithms we propose in this section enhance the techniques for homogenous CCP to heterogeneous CCP. All algorithms discussed in this section require an initial prefix-sum operation on the task-weight array $\mathcal{W}$ for the efficiency of subsequent subchain-weight computations. The prefix-sum operation replaces the ith entry $\mathcal{W}[i]$ with the sum of the first i entries ($\sum_{h=1}^{i} w_h$) so that computational weight W_{ij} of a task subchain $\mathcal{T}_{ij}$ can be efficiently determined as $\mathcal{W}[j] - \mathcal{W}[i-1]$ in $O(1)$ time. In our discussions, $\mathcal{W}$ is used to refer to the prefix-summed $\mathcal{W}$ array, and $O(N)$ cost of this initial prefix-sum operation is considered in the complexity analysis. Similarly, $E_{a,b}$ can be computed in $O(1)$ time on a prefix-summed processor-speed array. In all algorithms, we focus only on finding the optimal solution value, since an optimal solution can be easily constructed, once the optimal solution value is known.

Unless otherwise stated, *BINSEARCH* represents a binary search that finds the index to the element that is closest to the target value. There are variants of *BINSEARCH* to find the index of the greatest element not greater than the target value, and we will state whenever such variants are needed. *BINSEARCH* takes four parameters: the array to search, the start and end indices of the sub-array, and the target value. The range parameters are optional, and their absence means that the search will be performed on the whole array.

4.1 Heuristics

We propose a heuristic, *RB*, based on the recursive bisection idea. During each bisection, *RB* performs a two step process. First, it divides the current

```
RB (W, E, p, r)
if p = r then
   return;
Wtot ← W_{s_{p-1}+1,s_r};
q ← (p + r − 1)/2;
Wfirst ← Wtot × E_{p,q}/E_{p,r};
W ← Wfirst + W_{1,s_{p-1}};
s_q ←BINSEARCH(W, s_{p-1}, s_r, W);
RB(W, E, p, q);
RB(W, E, q + 1, r);
```

```
MP (W, N, E, P)
for p ← 1 to P do
   w ← W_{1,N} × E_{1,p}/E_{1,P};
   s_p ←BINSRCH(W, w, s_{p-1}, N);
```

Fig. 1. Heterogeneous CCP heuristics

processor chain $\mathcal{P}_{p,r}$ into two subchains $\mathcal{P}_{p,q}$ and $\mathcal{P}_{q+1,r}$. Then, it divides the current task chain $\mathcal{T}_{h,j}$ into two subchains $\mathcal{T}_{h,i}$ and $\mathcal{T}_{i+1,j}$ in proportion to the computational powers of the respective processor subchains. That is, the task separator index i is chosen such that the ratio $W_{h,i}/W_{i+1,j}$ is close to the ratio $E_{p,q}/E_{q+1,r}$, as much as possible. RB achieves optimal bisections at each level, however, the quality of the overall partition may be far away from that of the optimal solution.

We have investigated two metrics for bisecting the processor chain: chain length and chain processing power. The chain length metric divides the current processor chain $\mathcal{P}_{p,r}$ into two equal-*length* processor subchains, whereas the chain processing power metric divides $\mathcal{P}_{p,r}$ into two equal-*power* subchains. Since the first metric performed slightly better than the latter in our experiments, we will only discuss the chain length metric here. The pseudocode of the RB algorithm is given in Fig. 1, where the initial invocation takes its parameters as $(\mathcal{W}, \mathcal{E}, 1, P)$ with $s_0 = 0$ and $s_P = N$. Note that s_{p-1} and s_r are already determined at higher levels of recursion. *Wtot* is the total weight of current task subchain, and *Wfirst* is the weight for the first processor subchain in proportion to its processing speed. We need to add $W_{1,s_{p-1}}$ to *Wfirst* to seek it in the prefix-summed $\mathcal{W}$ array.

We also propose a generalization of Miguet and Pierson's heuristic, *MP* [11]. *MP* computes the separator index of each processor by considering that processor as a division point for the whole processor chain. In our version, the load assigned to the processor chain $\mathcal{P}_{1,p}$ is set to be proportional to the computational power $E_{1,p}$ of this subchain, as shown in Fig. 1.

Both *RB* and *MP* can be implemented in $O(N+P \lg N)$ time, where the $O(N)$ time is due to prefix-sum operation on the task-weight array.

Below, we investigate the theoretical bounds on the quality of these two heuristics. We assume P is a power of 2 for simplicity.

Lemma 4.1 *B_{RB} is upper bounded by $B^*+w_{\max}/e_{\min} - w_{\max}/(Pe_{\min})$.*

Proof: We use induction, and the basis is easy to show for $P = 2$. For the inductive step, assume the hypothesis holds for any number of processors less than P. Consider the first bisection, where the processors are split into two subchains, each containing $P/2$ processors. Let the total processing power in the left subchain be E_{left}. *RB* will distribute the workload array between the left and right processor subchains as evenly as possible. There will be a task t_i such that the left processor subchain will weigh more than the right subchain if t_i is assigned to the left subchain, and vice versa. Without loss of generality, assume that t_i is assigned to the left subchain. In the worst case, t_i is the maximum weighted task, and the total task weight assigned to the left subchain, W_{left}, can be upper bounded by

$$W_{\text{left}} \leq \frac{(W_{\text{tot}} + w_{\max})E_{\text{left}}}{E_{\text{tot}}}.$$

Using the inductive hypothesis, the bottleneck value among the processors of the left processor subchain can be upper bounded as follows.

$$
\begin{aligned}
B_{RB} \leq \frac{W_{\text{left}}}{E_{\text{left}}} + \frac{w_{\max}}{e_{\min}} - \frac{w_{\max}}{e_{\min} P/2} &\leq \frac{W_{\text{tot}} + w_{\max}}{E_{\text{tot}}} + \frac{w_{\max}}{e_{\min}} - \frac{w_{\max}}{e_{\min} P/2} \\
= B^* + \frac{w_{\max}}{E_{\text{tot}}} + \frac{w_{\max}}{e_{\min}} - \frac{w_{\max}}{e_{\min} P/2} &\leq B^* + \frac{w_{\max}}{e_{\min} P} + \frac{w_{\max}}{e_{\min}} - \frac{w_{\max}}{e_{\min} P/2} \\
= B^* + \frac{w_{\max}}{e_{\min}} - \frac{w_{\max}}{P e_{\min}}
\end{aligned}
$$

The same bound applies to the right processor subchain directly by the inductive hypothesis, since the right processor subchain is already underloaded.

■

Lemma 4.2 *B_{MP} is upper bounded by $B^* + w_{\max}/e_{\min}$.*

Proof: Let the sequence $\langle s_0, s_1, \ldots, s_P \rangle$ be the partition constructed by *MP*. For a processor $\mathcal{P}_p$, s_p is chosen to be the separator that best divides $\mathcal{P}_{1,p}$ and $\mathcal{P}_{p+1,P}$. Based on our discussion of bipartitioning quality in the proof of Lemma 4.1, W_{1,s_p} is bounded by

$$
E_{1,p} B^* - \frac{w_{\max}}{2} \leq W_{1,s_p} \leq E_{1,p} B^* + \frac{w_{\max}}{2}
$$

So, the load of processor p is upper bounded by

$$
\begin{aligned}
\frac{W_{1,s_p} - W_{1,s_{p-1}}}{e_p} &\leq \frac{E_{1,p} B^* + w_{\max}/2 - E_{1,p-1} B^* + w_{\max}/2}{e_p} \\
&= B^* + \frac{w_{\max}}{e_p} \leq B^* + \frac{w_{\max}}{e_{\min}}
\end{aligned}
$$

■

4.2 Dynamic Programming

The overlapping subproblems and the optimal substructure properties of the homogenous CCP can be extended to the heterogeneous CCP, and thus enabling dynamic programming solutions. The recursive definition for the bot-

tleneck value of an optimal partition can be derived as

$$B_i^p = \min_{0 \le j \le i} \left\{ \max \left\{ B_j^{p-1}, \frac{W_{j+1,i}}{e_p} \right\} \right\} \tag{4}$$

for the heterogeneous case. As in the homogenous case, B_i^p denotes the optimal solution value for partitioning the first i tasks onto the first p processors. This definition results in an $O(N^2P)$-time DP algorithm.

We generalize the observations of Choi and Narahari [3] to develop an $O(NP)$-time algorithm for heterogeneous systems as follows. Their first observation relies on the fact that the optimal position of the separator for partitioning the first i tasks cannot be to the left of the optimal position for the first $i-1$ tasks, i.e., $j_i^p \ge j_{i-1}^p$. Their second observation is that we need to advance a separator index only when the last part is overloaded and can stop when this is no longer the case, i.e., $B_j^{p-1} \ge W_{j+1,i}/e_p$. Then an optimal j_i^p can be chosen to correspond to the minimum of $\max\{B_j^{p-1}, W_{j+1,i}/e_p\}$ and $\max\{B_{j-1}^{p-1}, W_{j,i}/e_p\}$. That is, the recursive definition becomes:

$$B_i^p = \max \left\{ B_{j_i^p}^{p-1}, \frac{W_{j_i^p+1,i}}{e_p} \right\}, \text{ where } j_i^p = \operatorname*{argmin}_{j_{i-1}^p \le j \le i} \left\{ \max \left\{ B_j^{p-1}, \frac{W_{j+1,i}}{e_p} \right\} \right\}.$$

It is clear that the search ranges of separators overlap at only one position, and thus we can compute all B_i^p entries for $1 \le i \le N$ in only one pass over the task subchain. This reduces the complexity of the algorithm to $O(NP)$. Fig. 2(a) presents this algorithm.

Olstad and Manne reduced the complexity further to $O((N–P)P)$ by observing that there is no merit in leaving a processor empty, and thus the search for j_i^p can start at p instead of 1. However, this does not apply to the heterogeneous CCP, since it might be beneficial to leave a processor empty.

```
DP (W, N, P, E)
for i ← 1 to N do
  B[1, i] ← W_{1,i}/e_1;
for p ← 2 to P do
  j ← 0;
  for i ← j + 1 to N do
    if W_{j+1,i}/e_p ≤ B[p − 1, j] then
      B[p, i] ← B[p − 1, j];
    else
      repeat
        j ← j + 1;
      until W_{j+1,i}/e_p ≤ B[p − 1, j] or j ≥ i;
      if W_{j,i}/e_p < B[p − 1, j] then
        j ← j − 1;
        B[p, i] ← W_{j+1,i}/e_p;
      else
        B[p, i] ← B[p − 1, j];
return B_opt ← B[P, N];
```

(a)

```
DP+ (W, N, E, P, SL, SH)
for i ← SL_1 to SH_1 do
  B[1, i] ← W_{1,i}/e_1;
for p ← 2 to P do
  j ← SL_{p−1};
  for i ← SL_p to SH_p do
    if W_{j+1,i}/e_p ≤ B[p − 1, j] then
      B[p, i] ← B[p − 1, j];
    else
      repeat
        j ← j + 1;
      until W_{j+1,i}/e_p ≤ B[p − 1, j] or j ≥ i;
      if W_{j+1,i}/e_p = B[p − 1, j] then
        B[p, i] ← B[p − 1, j];
      else if W_{j,i}/e_p < B[p − 1, j] then
        j ← j − 1;
        B[p, i] ← W_{j+1,i}/e_p;
      else
        B[p, i] ← B[p − 1, j];
return B[P, N];
```

(b)

Fig. 2. DP algorithms for heterogeneous systems: (a) basic DP algorithm, and (b) DP algorithm (*DP+*) with static separator index bounding.

We propose another DP algorithm by extending the *DP+* algorithm (*DP* algorithm with static separator-index bounding) of Pınar and Aykanat [15] for the heterogeneous case. *DP+* limits the search space of each separator to avoid redundant calculation of B_i^p values. *DP+* achieves this separator index bounding by running left-to-right and right-to-left probe functions with the upper and lower bounds on the optimal bottleneck value.

We extend the probing operation to the heterogeneous case as shown in Fig. 3. In the figure, *LR-PROBE* and *RL-PROBE* denote the left-to-right probe and right-to-left probe, respectively. These algorithms not only decide whether a candidate value is a feasible bottleneck value, but they also set the separator index (s_p) values for their greedy approach. In *LR-PROBE*, *BIN-*

```
LR-PROBE (W, N, E, P, B)
sum ← 0;
for p ← 1 to P − 1 do
  myB ← B × e_p;
  Bsum ← sum + myB;
  m ← BINSEARCH(W, Bsum);
  sum ← W_{1,m};
  s_p ← m;
if sum + B × e_P ≥ W_{1,N} then
  return TRUE;
else
  return FALSE;
```

(a)

```
RL-PROBE (W, N, E, P, B)
sum ← W_{1,N};
for p ← P downto 2 do
  myB ← B × e_p;
  Bsum ← sum − myB;
  m ← BINSEARCH(W, Bsum);
  sum ← W_{1,m};
  s_{p−1} ← m;
if sum − B × e_1 ≤ 0 then
  return TRUE;
else
  return FALSE;
```

(b)

Fig. 3. Greedy *PROBE* algorithms for heterogeneous systems: (a) left-to-right, and (b) right-to-left.

SEARCH$(\mathcal{W}, w)$ searches $\mathcal{W}$ for the largest index m such that $W_{1,m} \leq w$. Similarly, in *RL-PROBE*, *BINSEARCH*$(\mathcal{W}, w)$ refers to a binary search algorithm that searches $\mathcal{W}$ for the smallest index m such that $W_{1,m} \geq w$.

DP+, as presented in Fig. 2(b), uses Lemma 4.3 to limit the search space of s_p values.

Lemma 4.3 *For a given heterogeneous CCP instance* $(\mathcal{W}, N, \mathcal{E}, P)$*, a feasible bottleneck value* UB *and a lower bound on the bottleneck value* LB*; let the sequences* $\Pi^1 = \langle h_0^1, h_1^1, \ldots, h_P^1 \rangle$*,* $\Pi^2 = \langle l_0^2, l_1^2, \ldots, l_P^2 \rangle$*,* $\Pi^3 = \langle l_0^3, l_1^3, \ldots, l_P^3 \rangle$ *and* $\Pi^4 = \langle h_0^4, h_1^4, \ldots, h_P^4 \rangle$ *be the partitions constructed by LR-PROBE*(UB)*, RL-PROBE*(UB)*, LR-PROBE*(LB) *and RL-PROBE*(LB)*, respectively. Then, an optimal partition* $\Pi_{\text{opt}} = \langle s_0, s_1, \ldots, s_P \rangle$ *satisfies* $SL_p \leq s_p \leq SH_p$ *for all* $1 \leq p \leq P$*, where* $SL_p = \max\{l_p^2, l_p^3\}$ *and* $SH_p = \min\{h_p^1, h_p^4\}$*.*

Proof: We know that any feasible bottleneck value is greater than or equal to the optimal bottleneck value, i.e., $UB \geq B_{\text{opt}}$. Consider h_p^1, which is the largest index such that the first h_p^1 tasks can be partitioned over p processors without exceeding UB. Then $s_p > h_p^1$ implies $B_{\text{opt}} > UB$, a contradiction. So,

$s_p \leq h_p^1$. Since, *RL-PROBE* is just the symmetric algorithm of *LR-PROBE*, the same argument proves $s_p \geq l_p^2$.

Consider the optimal partition constructed by *RL-PROBE*(B_{opt}). Since $B_{\text{opt}} \geq LB$, by the greedy property of *RL-PROBE*, $s_p \leq h_p^4$. Assume $s_p < l_p^3$ for some p, then another partition obtained by advancing the s_p value to l_p^3 does not increase the bottleneck value, since the first l_p^3 tasks are successfully partitioned over the first p processors without exceeding LB and thus B_{opt}. An optimal partition $\Pi_{\text{opt}} = \langle s_0, s_1, \ldots, s_P \rangle$ satisfies $l_p^3 \leq s_p \leq h_p^4$. ■

The lower bound LB can be initialized to the optimal lower bound when all processors are equally loaded as

$$LB = B^* = \frac{W_{\text{tot}}}{E_{\text{tot}}}. \tag{5}$$

An upper bound UB can be computed in practice with a fast and effective heuristic, and Lemma 4.1 provides a theoretically robust bound as

$$UB = B^* + \frac{w_{\max}}{e_{\min}} - \frac{w_{\max}}{Pe_{\min}}. \tag{6}$$

4.3 Parametric Search

Parametric search algorithms can be constructed with a *PROBE* function (either LR-PROBE or RL-PROBE given in Fig. 3), and a method to search the space of candidate values. Below, we describe several algorithms to search the space of bottleneck values for the heterogeneous case.

4.3.1 Nicol's Algorithm

We revise Nicol's algorithms for heterogeneous systems as follows. The candidate B values become task subchain weights divided by processor subchain speeds. The algorithm starts with searching for the smallest j so that

```
NICOL (W, E, N, P)
i_0 ← 1;
for b ← 1 to P − 1 do
  ilow ← i_{b−1};  ihigh ← N;
  while ilow < ihigh do
    imid ← (ilow + ihigh)/2;
    B ← W_{i_{b−1},imid}/e_b;
    if PROBE(B) then
      ihigh ← imid;
    else
      ilow ← imid + 1;
  i_b ← ihigh;
  B_b ← W_{i_{b−1},i_b}/e_b;
B_P ← W_{i_{P−1},N}/e_P;
return B_opt ← min_{1≤b≤P}{B_p};
```

```
NICOL+ (W, E, N, P)
i_0 ← 1;
LB ← B* ← W_{1,N}/E_{1,P};
UB ← LB + w_max × (1/e_min − 1/E_tot);
for b ← 1 to P − 1 do
  ilow ← i_{b−1};  ihigh ← N;
  while ilow < ihigh do
    imid ← (ilow + ihigh)/2;
    B ← W_{i_{b−1},imid}/e_b;
    if LB ≤ B < UB then
      if PROBE(B) then
        ihigh ← imid;
        UB ← B;
      else
        ilow ← imid + 1;
        LB ← B;
    else if B ≥ UB then
      ihigh ← imid;
    else
      ilow ← imid + 1;
  i_b ← ihigh;
  B_b ← W_{i_{b−1},i_b}/e_b;
B_P ← W_{i_{P−1},N}/e_P;
return B_opt ← min_{1≤b≤P}{B_p};
```

Fig. 4. Nicol's algorithms for heterogeneous systems: (a) Nicol's basic algorithm, (b) Nicol's algorithm (*NICOL+*) with dynamic bottleneck-value bounding.

probing with $W_{1,j}/e_1$ succeeds, and probing with $W_{1,j-1}/e_1$ fails. This means $W_{1,j-1}/e_1 < B_{\text{opt}} \leq W_{1,j}/e_1$, and thus in an optimal solution the probe function will assign the first j tasks to the first processor if it is the bottleneck processor, and the first $j-1$ tasks to the first processor if not. Then the optimal solution value is the minimum of $W_{1,j}/e_1$ and the optimal solution value for partitioning the remaining subchain $\mathcal{T}_{j,N}$ to $P-1$ processors, since any solution with a bottleneck value less than $W_{1,j}/e_1$ will assign only the first $j-1$ tasks to the first processor. Finding the j value requires $\lg N$ probes, and we repeat this search operation for all processors in order. This version of Nicol's algorithm runs in $O(N+(P \lg N)^2)$ time. Fig. 4(a) displays this algorithm.

By keeping the largest B that succeeded and the smallest B that failed, we can improve Nicol's algorithm by eliminating unnecessary probing. Let LB and UB represent the lower bound and upper bound for B_{opt}, respectively. If a processor cannot update LB or UB, that processor does not make any *PROBE* calls. This algorithm, presented in Fig. 4(b), is referred to as *NICOL+*.

In the worst case, a processor makes $O(\lg N)$ *PROBE* calls. But, as we will prove below, the number of probes performed by *NICOL+* cannot exceed $P \lg (1+w_{\max}/(Pe_{\min}w_{\min}))$. This analysis also improves known complexities of homogeneous version of the algorithm. Lemma 4.4 describes an upper bound on the number of probes performed by *NICOL+* algorithm.

Lemma 4.4 *The number of probes required by NICOL+ is upper bounded by* $P \lg (1+(UB - LB) / (Pw_{\min}))$.

Proof: Consider the first step of the algorithm, where we search for the smallest separator index that makes the first processor the bottleneck processor. We can restrict this search in a range that covers only those indices for which the weight of the first chain will be in the $[LB, UB]$ interval. If there are n_1 tasks in this range, *NICOL+* will require $\lg n_1$ probes. This means that the $[LB, UB]$ interval is narrowed by at least $(n_1 - 1)w_{\min}$ after the first step.

Let k_p be the number of probes by the pth processor. Since k_p probes narrows the $[LB, UB]$ interval by $\left(2^{k_p} - 1\right) w_{\min}$, we have

$$\left(\left(2^{k_1} - 1\right) + \left(2^{k_2} - 1\right) + \ldots + \left(2^{k_{P-1}} - 1\right)\right) w_{min} \leq UB - LB,$$

and thus $2^{k_1} + 2^{k_2} + \ldots + 2^{k_{P-1}} \leq \dfrac{UB - LB}{w_{\min}} + P - 1$. The corresponding total

number of probes is $\sum_{p=1}^{P-1} k_p$, which reaches its maximum when $\sum_{p=1}^{P-1} 2^{k_p}$ is maximum and $k_1 = k_2 = \ldots = k_{P-1} = k$ for some k. In that case,

$$(P-1)2^k \leq \frac{UB - LB}{w_{\min}} + P - 1$$

and thus

$$k \leq \lg\left(1 + \frac{UB - LB}{w_{\min}(P-1)}\right).$$

So, the total number of probes performed by *NICOL+* is upper bounded by:

$$\sum_{p=1}^{P-1} k_p \leq (P-1)k \leq (P-1)\lg\left(1 + \frac{UB - LB}{w_{\min}(P-1)}\right) < P\lg\left(1 + \frac{UB - LB}{w_{\min}P}\right)$$

■

Corollary 4.5 *NICOL+ requires at most* $P\lg(1+w_{\max}/(Pe_{\min}w_{\min}))$ *probes for heterogeneous, and* $P\lg(1 + w_{\max}/(Pw_{\min}))$ *probes for homogeneous systems.*

NICOL+ runs in $O(N+P^2 \lg N \lg(1+w_{\max}/(Pe_{\min}w_{\min})))$ time, with the $O(P \lg N)$ cost of a *PROBE* call. In most configurations, $w_{\max}/(e_{\min}w_{\min}P)$ is very small, and is $O(1)$ if $Pe_{\min} = \Omega(w_{\max}/w_{\min})$. In that case, the runtime complexity of *NICOL+* reduces to $O(N+P^2 \lg N)$.

4.3.3 Bidding Algorithm

For heterogeneous systems, the bidding algorithm uses the lower bound given in Eq. 5 for optimal bottleneck value, and gradually increases this lower bound. The bid of each processor $\mathcal{P}_p$, for $p = 1, 2, \ldots, P-1$, is calculated as $W_{s_{p-1}+1,s_p+1} \,/\, e_p$, which is equal to the load of $\mathcal{P}_p$ if it also executes the first task of $\mathcal{P}_{p+1}$ in addition to its current load. Then, the algorithm selects the processor with the minimum bid value so that this bid value becomes the next bottleneck value to be considered for feasibility. The processors following

```
BIDDING (W, N, E, P)
minBid ← W_{1,N}/E_{1,P};
LR-PROBE(W, N, E, P, minBid);
for p ← 1 to P − 1 do
  bids[p] ← W_{s_{p−1}+1,s_p+1}/e_p;
Q ←BUILD-HEAP(P);
repeat
  minP ←EXTRACT-MIN(Q);
  wlast ← W_{s_{P−1}+1,N}/e_P;
  minBid ← bids[minP];
  if minBid < wlast then
    for p ← minP to P − 1 do
      s_p ← BINSEARCH(W, minBid × e_p + W_{1,s_{p−1}});
      previousBid ← bids[p];
      bids[p] ← W_{s_{p−1}+1,s_p}/e_p;
      if bids[p] > previousBid then
        INCREASE-KEY(Q, p);
      else if bids[p] < previousBid then
        DECREASE-KEY(Q, p);
until minBid ≥ wlast;
```

Fig. 5. Bidding algorithm for heterogeneous systems.

the bottleneck processor in the processor chain are processed in order, except the last processor. The separator indices of these processors are adjusted accordingly so that the processors are maximally loaded not to exceed that new bottleneck value. The load of the last processor determines the feasibility of the current bottleneck value. If current bottleneck value is not feasible, the process repeats. Fig. 5 presents the bidding algorithm, which uses a min-priority queue that maintains the processors keyed according to their bid values.

In the worst case, the bidding algorithm moves P separators for $O(N)$ positions. Choosing a new bottleneck value takes $O(\lg P)$ time using a binary heap implementation of the priority queue. Totally the complexity of the algorithm is $O(NP \lg P)$ in the worst case. Despite this high worst-case complexity, the bidding algorithm is quite fast in practice.

For heterogeneous systems, the bisection algorithm can use the LB and UB values given in Eqs. 5 and 6. A binary search on this $[LB, UB]$ interval requires $O(\lg(w_{\max}/(\epsilon E_{\text{tot}})))$ probes, thus leading to an $O(\lg(w_{\max}/(\epsilon E_{\text{tot}}))P \lg N)$-time algorithm, where ϵ is the specified accuracy of the algorithm. Fig. 6(a) presents this ϵ-approximation bisection algorithm. We should note that, although the homogenous version of this algorithm becomes an exact algorithm for integer-valued workload arrays by setting $\epsilon = 1$, this is not the case for heterogeneous systems.

We enhance this bisection algorithm to be an exact algorithm for heterogeneous systems by extending the scheme proposed by Pınar and Aykanat [15] for homogenous systems. After each probe, we move lower and upper bounds to realizable bottleneck values, as opposed to the probed value. In heterogeneous systems, realizable bottleneck values are subchain weights divided by appropriate processor speeds. After a successful probe, we decrease UB to the bottleneck value of the partition constructed by the probe, and after a failed probe we increase LB to the bid value as described for the bidding algorithm in Section 4.3.3. Each probe eliminates at least one candidate bottleneck value, and thus the bisection algorithm terminates in a finite number of steps with an optimal solution. Fig. 6(b) displays the exact bisection algorithm.

5 Chain Partitioning (CP) Problem for Heterogeneous Systems

In this section, we study the problem of partitioning a chain of tasks onto a set of processors, as opposed to a chain of processors. The solution to

```
BISECTION (W, N, E, P, ε)
LB ← W_{1,N}/E_{1,P};
UB ← LB + w_max/e_min;
while UB − LB ≥ ε do
  midB ← (UB + LB)/2;
  if PROBE(midB) then
    UB ← midB;
  else
    LB ← midB;
return UB;
```

(a)

```
EXACT-BISECTION (W, N, E, P)
LB ← W_{1,N}/E_{1,P};
UB ← LB + w_max/e_min;
while UB > LB do
  midB ← (UB + LB)/2;
  if LR-PROBE(midB) then
    UB ← min_{1≤p≤P} W_{s_{p−1},s_p}/e_p;
  else
    LB ← min_{1≤p≤P−1} W_{s_{p−1},s_p+1}/e_p;
return UB;
```

(b)

Fig. 6. Bisection algorithms for heterogeneous systems: (a) ϵ-approximation bisection algorithm, (b) Exact bisection algorithm.

this problem is not only separators on the task chain, but also processor-to-subchain assignments. Thus, we define a mapping $\mathcal{M}$ as a partition $\Pi = \langle s_0 = 0, s_1, \ldots, s_P = N \rangle$ of the given task chain $\mathcal{T} = \langle t_1, t_2, \ldots t_N \rangle$ with $s_p \leq s_{p+1}$ for $0 \leq p < P$, and a permutation $\langle \pi_1, \pi_2, \ldots, \pi_P \rangle$ of the given set of P processors $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_P\}$. According to this mapping, the pth task subchain $\langle t_{s_{p-1}+1}, \ldots, t_{s_p} \rangle$ is executed on processor $\mathcal{P}_{\pi_p}$. The cost $C(\mathcal{M})$ of a mapping $\mathcal{M}$ is the maximum subchain computation time, determined by the subchain weight and the execution speed of the assigned processor, i.e.,

$$C(\mathcal{M}) = \max_{1 \leq p \leq P} \left\{ \frac{W_{s_{p-1}+1, s_p}}{e_{\pi_p}} \right\}.$$

We will prove that the CP problem is NP-complete. The decision problem for the CP problem for heterogeneous systems is as follows.

Given a chain of tasks $\mathcal{T} = \langle t_1, t_2, \ldots, t_N \rangle$, a weight $w_i \in Z^+$ for each $t_i \in \mathcal{T}$, a set of processors $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_P\}$ with $P < N$, an execution speed $e_p \in Z^+$ for each $\mathcal{P}_p \in \mathcal{P}$, and a bound B, decide if there exists a mapping $\mathcal{M}$ of $\mathcal{T}$ onto $\mathcal{P}$ such that $C(\mathcal{M}) \leq B$.

Theorem 5.1 *The CP problem for heterogeneous systems is NP-complete.*

Proof: We use reduction from the 3-Partition (3P) problem. A pseudo-polynomial transformation suffices, because 3P problem is NP-complete in the strong sense (i.e., there is no pseudo-polynomial time algorithm for the problem unless P=NP). The 3P problem is stated in [6] as follows.

Given a finite set $\mathcal{A}$ of $3m$ elements, a bound $B \in Z^+$, and a cost $c_i \in Z^+$ for each $a_i \in \mathcal{A}$, where $\sum_{a_i \in \mathcal{A}} c_i = mB$ and each c_i satisfies $B/4 < c_i < B/2$, decide if $\mathcal{A}$ can be partitioned into m disjoint sets $S_1, S_2, \ldots, S_m$ such that $\sum_{a_i \in S_p} c_i = B$ for $p = 1, 2, \ldots, m$.

For a given instance of the 3P problem, the corresponding CP problem is constructed as follows.

- The number of tasks N is $m(B+1)-1$. The weight of every $(B+1)$st task is B, (i.e., $w_i = B$ for $i \ mod \ (B+1) = 0$), and the weights of all other tasks are 1.
- The number of processors P is $4m-1$. The first $m-1$ processors have execution speeds of B, (i.e., $e_p = B$ for $p = 1, 2, \ldots, m-1$), and the remaining processors have execution speeds equal to the costs of items in the 3P problem (i.e., $e_p = c_{p-m+1}$ for $p = m, \ldots, 4m-1$).

We claim that there is a solution to the 3P problem if and only if there is a mapping $\mathcal{M}$ with cost $C(\mathcal{M}) = 1$ for the CP problem. The following observations constitute the basis for our proof.

- The processors with execution speeds of B must be mapped to tasks with weight B to have a solution with cost $C(\mathcal{M}) = 1$, because the execution speeds of all other processors are $\leq B/2$. These processors (tasks) serve as

divider processors (tasks).

- The total weight of the chain is $3m + (m-1)B = (B+3)m - B$. The sum of execution speeds of all processors is also $(m-1)B + 3m = (B+3)m - B$. This forces each processor to be assigned a load with value equal to its execution speed to achieve a mapping with cost $C(\mathcal{M}) = 1$.

As noted above, the divider processors should be assigned to the divider tasks. Between two successive divider tasks there is a subchain of B unit-weight tasks with total weight B, which must be assigned to a subset of processors with total execution speed B. Since there are m such subchains, the same grouping of the processors is also valid for grouping c_i values in the 3P problem. Thus the 3P problem can be reduced to the CP problem, proving the CP problem is NP-hard.

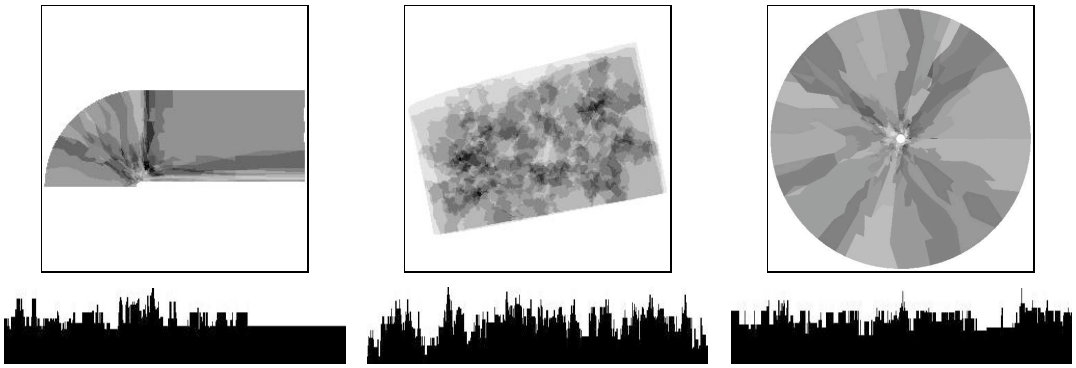
The cost of a given mapping can be computed in polynomial time, thus the problem is in NP. Thus we can conclude that the chain partitioning problem for heterogeneous systems is NP-Complete. ■

This complexity shows that we need to resort to heuristics for practical solutions to the CP problem. With the nearly perfect balance results and extremely fast runtimes as we will present in Section 6, CCP algorithms can serve as good heuristics for the CP problem. We tried this approach by finding optimal CCP solutions for randomly ordered processor chains of a CP instance. We observed that the sensitivity to processor ordering is quite low. You can find a description of these studies in Section 6.3. We also tried improvement techniques, where we swapped processors in the chain to decrease the bottleneck value, but the improvements were modest and could hardly compensate the increase in runtimes.

6 Experimental Results

6.1 Experimental Setup

The 1D task arrays used in both CCP and CP experiments were derived from two different applications: image-space-parallel direct volume rendering and row-parallel sparse matrix vector multiplication. Direct volume rendering experiments are performed on three curvilinear datasets from NASA Ames Research Center [1], namely *Blunt Fin* (blunt), *Combustion Chamber* (comb), and *Oxygen Post* (post). These datasets are processed using the tetrahedralization techniques described in [7] and [16] to produce three-dimensional (3D) unstructured volumetric datasets. The two-dimensional (2D) workload arrays are constructed by projecting 3D volumetric datasets onto 2D screens of resolution 256×256 using the workload criteria of image-space-parallel direct volume rendering algorithm described in [2]. Here, the rendering operations associated with the individual pixels of the screen constitute the computational tasks of the application. The resulting 2D task array is then mapped to a 1D task array using Hilbert space-filling-curve traversal [13]. The workload distributions of the 2D task arrays are visualized in Fig. 7, where darker areas represent more weighted tasks. The histograms at the bottom of the 2D pictures show the weight distributions of the resulting 1D task arrays. In the sparse matrix experiments, we consider rowwise block partitioning of the matrices obtained from University of Florida Sparse Matrix Collection [4]. In row-parallel matrix vector multiplies, the rows correspond to the tasks to be partitioned, and the number of nonzeros in each row is the weight of the corresponding task. The nonzero distributions of the sparse matrices are shown in Fig. 8. The histograms on the right side of the visualizations represent the number of nonzeros in each row.



a) Blunt Fin b) Combustion Chamber c) Oxygen Post

Fig. 7. Visualization of direct volume rendering dataset workloads. Top: workload distributions of 2D task arrays. Bottom: histogram showing weight distributions of 1D task chains.

Table 2 displays the properties of the 1D task chains used in our experiments. In the volume rendering dataset, the number of tasks is considerably less than the screen resolution, because zero-weight tasks are omitted. In the sparse matrix dataset, the number of tasks is equal to the number of rows.

We have experimented with $P = 32, 64, 128, 256, 512, 1024$, and 2048-way partitioning of each test data. In these experiments, the processor speeds are chosen uniformly distributed in the 1–20 range, i.e., $1 \leq e_p \leq 20$, for each processor $\mathcal{P}_p$. The P-way partitioning of a given task chain constitutes a partitioning instance.

In the experiments, the solution qualities are represented by percent load imbalance values. The percent load imbalance of a partition is computed as $100 \times (B - B^*)/B^*$, where B denotes the bottleneck value of the respective partition.

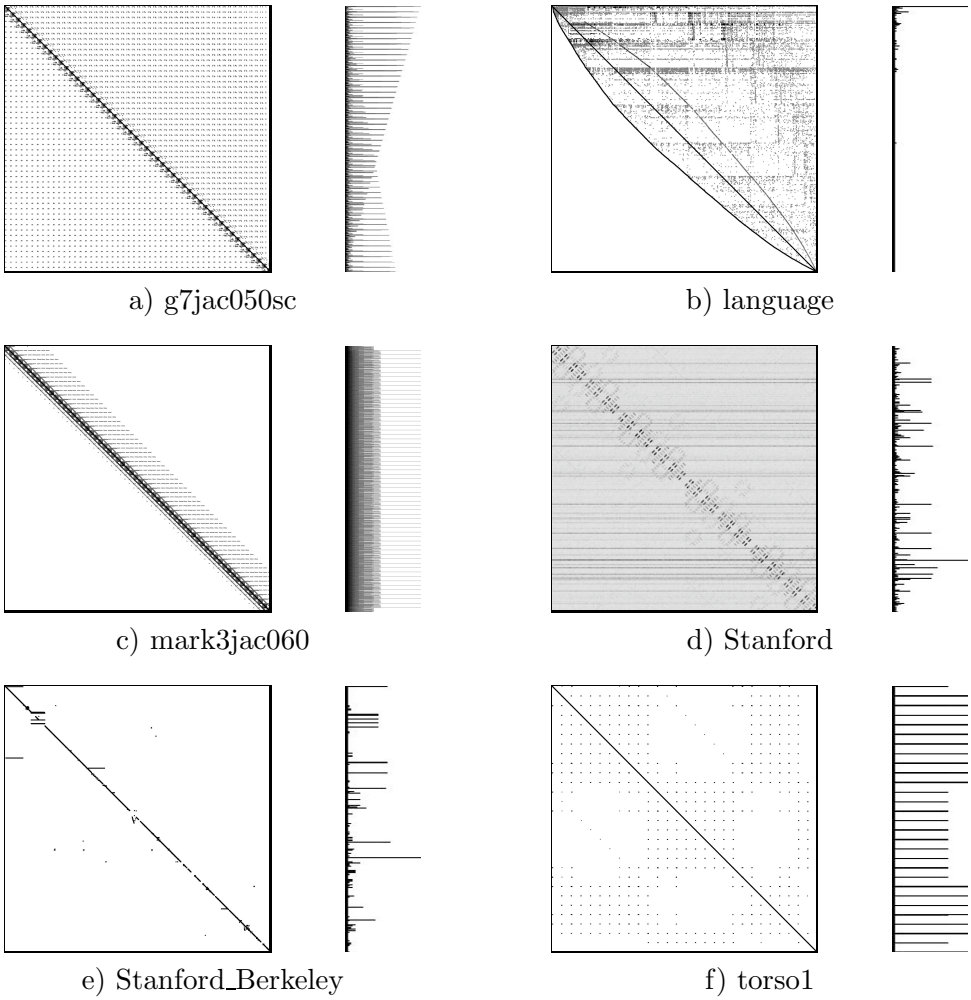


Fig. 8. Visualization of sparse matrix dataset workloads. Left: non-zero distribution of the sparse matrices. Right: histogram showing weight distributions of the 1D task chains.

6.2 CCP Experiments

The proposed CCP algorithms were implemented in the Java language. Tables 3 and 4 compare the solution qualities of heuristics with respect to those of the optimal partitions obtained by the exact algorithms. In these tables, OPT values refer to the optimal solution value. The bottom parts of these two tables show the geometric averages of the percent load imbalance values over number of processors. As seen in Tables 3 and 4, RB performs considerably

Table 2
Properties of the test set

Name	No. of tasks N	Workload			
		Total	Per task		
		W_{tot}	w_{avg}	w_{min}	w_{max}
Volume rendering dataset					
blunt	20.6 K	1.9 M	90.95	36	171
comb	32.2 K	2.1 M	64.58	14	149
post	49.0 K	5.4 M	109.73	33	199
Sparse matrix dataset					
g7jac050sc	14.7 K	0.2 M	10.70	2	149
language	399.1 K	1.2 M	3.05	1	11555
mark3jac060	27.4 K	0.2 M	6.22	2	44
Stanford	261.6 K	2.3 M	8.84	1	38606
Stanford_Berkeley	615.4 K	7.6 M	12.32	1	83448
torso1	116.2 K	8.5 M	73.32	9	3263

better than MP. Out of 63 partitioning instances, the RB and MP heuristics find the best solutions in 54 and 17 partitioning instances, respectively.

As seen in Tables 3 and 4, the quality gap between exact algorithms and heuristics increases with increasing number of processors. For instance, in 2048-way partitioning of the `torso1` matrix, best heuristic finds a solution with 186.67% load imbalance, which means a processor is loaded more than 2.8 times the average load. This will cause a slowdown as the number of processors increase. An optimal solution however, will have a load imbalance value of 28.43%, providing scalability to thousands of processors.

Tables 5 and 6 display the execution times of the proposed CCP algorithms on a workstation equipped with a 3 GHz Pentium-IV and 1 GB of memory. In these tables, DP+, NC+, BID and EBS respectively represent the *DP+*, *NICOL+*, *BIDDING* and *EXACT-BISECTION* algorithms presented in Figs. 2, 4, 5 and 6. For a better relative performance comparison, execution times of the algorithms are normalized with respect to those of the *RB* heuristic and averages of these normalized values over P are presented at the bottom of the

Table 3
Percent load imbalance values for the volume rendering dataset

CCP instance		Heuristics		OPT
Name	P	RB	MP	
blunt	32	0.13	0.21	0.07
	64	0.76	0.76	0.14
	128	1.52	1.52	0.32
	256	8.35	8.35	0.68
	512	10.66	25.31	1.21
	1024	30.85	50.19	2.40
	2048	63.08	92.52	5.01
comb	32	0.12	0.37	0.06
	64	0.85	1.23	0.13
	128	1.71	3.01	0.22
	256	4.57	4.57	0.43
	512	13.35	21.63	0.93
	1024	27.55	29.08	1.81
	2048	36.64	87.62	3.67
post	32	0.08	0.08	0.04
	64	0.37	0.37	0.07
	128	1.43	1.23	0.13
	256	2.04	2.04	0.26
	512	5.24	10.66	0.52
	1024	15.87	13.09	1.05
	2048	29.54	37.44	2.19
Geometric averages over P				
	32	0.11	0.18	0.05
	64	0.62	0.70	0.11
	128	1.55	1.78	0.21
	256	4.27	4.27	0.42
	512	9.07	18.00	0.84
	1024	23.80	26.73	1.66
	2048	40.87	67.20	3.43

table.

In Tables 5 and 6, relative performance comparison of heuristics shows that MP is slightly faster than RB. Since RB outperforms MP in terms of solution quality as shown in Tables 3 and 4, these results reveal the superiority of RB to MP.

In Tables 5 and 6, relative performances of exact CCP algorithms shows that both NICOL+ and EBS are an order of magnitude faster than DP+ and BID for both volume rendering and sparse matrix datasets. As also seen in these two tables, EBS is slightly faster than NICOL+.

It is worth highlighting that for small to medium concurrency, the time it takes EBS and NICOL+ algorithms to find optimal solutions is less than three times

Table 4
Percent load imbalance values for the sparse matrix dataset

CCP instance		Heuristics		OPT
Name	P	RB	MP	
g7jac050sc	32	1.29	2.72	0.33
	64	23.92	23.92	0.75
	128	10.03	12.95	1.59
	256	55.44	136.95	2.93
	512	121.67	203.52	6.86
	1024	141.33	804.56	13.96
	2048	540.52	719.69	29.58
language	32	1.03	0.93	0.07
	64	0.23	0.35	0.07
	128	1.19	7.30	0.13
	256	48.43	48.43	26.11
	512	169.40	169.40	155.93
	1024	433.24	1,347.37	406.58
	2048	1,160.51	1,160.51	908.41
mark3jac060	32	0.09	1.68	0.09
	64	1.04	1.04	0.19
	128	5.70	5.70	0.39
	256	10.43	10.43	0.63
	512	26.26	42.04	1.36
	1024	49.95	68.70	2.81
	2048	111.44	447.25	5.72
Stanford	32	21.32	38.49	3.68
	64	43.24	54.53	7.68
	128	67.09	116.07	17.73
	256	171.60	379.11	121.62
	512	499.67	1,844.58	349.75
	1024	1,086.98	3,749.09	790.23
	2048	1,765.39	2,451.80	1,672.12
Stanford_Berkeley	32	8.88	8.87	0.98
	64	90.20	92.58	2.49
	128	78.69	78.69	5.68
	256	120.40	1,123.38	46.08
	512	396.54	396.54	196.45
	1024	507.94	2,357.13	486.79
	2048	1,068.08	2,768.97	1,068.08
torso1	32	1.47	0.87	0.42
	64	2.02	6.64	0.81
	128	6.89	6.89	1.19
	256	43.40	43.65	1.89
	512	106.44	106.44	7.45
	1024	188.02	308.62	15.41
	2048	186.67	713.40	28.43
Geometric averages over P				
	32	1.80	3.28	0.38
	64	5.97	8.15	0.74
	128	11.64	17.99	1.46
	256	54.12	104.29	8.95
	512	150.05	219.83	30.10
	1024	261.19	766.69	67.49
	2048	539.31	1,103.86	140.83

the time of the fastest heuristic. More precisely, on average, EBS takes only 172.6% more time than the fastest heuristic for 256-way partitioning. On the other hand, at higher number of processors, the solution qualities of heuristics degrade significantly: on average, optimal solutions provide 9.44, 9.35 and

Table 5
Partitioning times (in msecs) for the volume rendering dataset

CCP Instance		Heuristics			Exact Algorithms			
Name	P	*RB*1	*RB*2	*MP*	DP+	NC+	BID	EBS
blunt	32	0.34	0.34	0.33	1	0.48	0.50	0.44
	64	0.35	0.37	0.35	1	0.70	0.73	0.60
	128	0.39	0.42	0.37	2	1.21	1.96	0.86
	256	0.45	0.51	0.42	5	1.95	5.63	1.71
	512	0.59	0.71	0.52	14	2.72	13.85	3.24
	1024	0.83	1.10	0.71	54	8.24	40.91	6.04
	2048	1.32	1.93	1.06	213	11.59	125.67	13.39
comb	32	0.52	0.53	0.52	1	0.67	0.71	0.64
	64	0.54	0.55	0.53	1	0.82	1.00	0.80
	128	0.57	0.60	0.56	2	1.39	2.17	1.26
	256	0.64	0.70	0.61	5	2.21	5.72	2.08
	512	0.78	0.91	0.71	17	4.00	19.16	3.59
	1024	1.03	1.31	0.90	62	7.27	58.37	7.05
	2048	1.53	2.16	1.31	242	13.70	174.19	13.05
post	32	0.81	0.82	0.81	2	0.96	0.96	0.91
	64	0.83	0.84	0.82	2	1.22	1.44	1.10
	128	0.86	0.89	0.85	3	1.74	2.67	1.52
	256	0.93	1.00	0.90	5	2.72	6.48	2.44
	512	1.08	1.20	1.01	16	4.55	17.65	3.70
	1024	1.34	1.61	1.21	55	8.89	55.77	7.52
	2048	1.86	2.46	1.63	216	11.12	148.28	13.88
Averages normalized w.r.t. RB1 times								
	32	1.00	1.01	0.99	2	1.29	1.34	1.22
	64	1.00	1.03	0.99	2	1.66	1.88	1.50
	128	1.00	1.05	0.97	4	2.51	3.98	2.06
	256	1.00	1.10	0.95	8	3.57	9.46	3.22
	512	1.00	1.16	0.91	21	4.67	21.52	4.52
	1024	1.00	1.26	0.87	55	7.85	49.04	6.56
	2048	1.00	1.40	0.85	145	7.90	96.18	8.70

8.30 times better load imbalance values than the best heuristic for 512, 1024 and 2048-way partitionings, respectively. According the these experimental results, we recommend the use of exact CCP algorithms instead of heuristics for heterogeneous systems.

6.3 CP Experiments

Tables 7 and 8 display the results of our experiments to show the sensitivity of the processor orderings on the solution quality of CP problem instances. In these experiments, we find the optimal CCP solutions for R randomly ordered processor chains of a CP instance, and display geometric averages of the best and average load imbalance values over number of processors. As seen in the tables, for a fixed P, the average imbalance values almost remain the same, as

Table 6
Partitioning times (in msecs) for the sparse matrix dataset

CCP Instance		Heuristics			Exact Algorithms			
Name	P	*RB*1	*RB*2	*MP*	DP+	NC+	BID	EBS
g7jac050sc	32	0.25	0.26	0.25	1	0.48	0.50	0.40
	64	0.26	0.28	0.26	1	0.66	1.12	0.61
	128	0.30	0.33	0.28	4	1.00	3.07	0.93
	256	0.36	0.42	0.33	13	2.19	8.11	1.78
	512	0.49	0.62	0.42	58	3.70	27.24	3.22
	1024	0.72	1.01	0.60	228	5.23	76.64	6.75
	2048	1.19	1.80	0.95	1744	15.62	206.17	12.12
language	32	8.46	8.47	8.46	18	8.68	10.01	8.67
	64	8.48	8.50	8.47	20	9.11	10.79	8.93
	128	8.53	8.55	8.51	27	9.81	14.29	9.67
	256	8.61	8.67	8.58	1484	11.80	8.80	11.03
	512	8.81	8.94	8.73	5576	11.20	9.13	11.96
	1024	9.31	9.60	9.16	16625	14.94	9.72	16.19
	2048	10.14	10.75	9.85	33115	17.24	10.89	21.22
mark3jac060	32	0.45	0.46	0.45	1	0.65	0.59	0.59
	64	0.47	0.48	0.46	1	0.86	0.91	0.71
	128	0.50	0.53	0.49	3	1.06	1.81	1.10
	256	0.57	0.62	0.54	7	2.27	4.57	1.75
	512	0.70	0.82	0.64	44	3.22	10.68	3.44
	1024	0.95	1.21	0.83	176	7.81	27.67	6.28
	2048	1.44	1.99	1.20	782	14.26	83.03	10.59
Stanford	32	7.63	7.65	7.63	37	7.93	46.99	7.95
	64	7.65	7.67	7.65	119	8.57	215.40	8.27
	128	7.69	7.72	7.68	1590	9.38	709.72	8.93
	256	7.77	7.84	7.74	7769	9.79	6983.20	10.21
	512	7.95	8.08	7.87	18887	11.06	17990.28	12.52
	1024	8.35	8.65	8.20	45441	13.50	40059.93	17.34
	2048	9.16	9.77	8.82	95157	21.96	49598.76	24.69
Stanford_Berkeley	32	13.95	13.96	13.95	46	14.33	35.83	14.31
	64	13.98	13.99	13.97	119	14.99	111.68	14.74
	128	14.02	14.05	14.00	505	16.12	401.75	15.61
	256	14.10	14.17	14.07	4832	18.81	3426.23	17.19
	512	14.28	14.42	14.21	20308	21.84	12252.13	19.20
	1024	14.73	15.01	14.57	49328	26.96	26545.32	23.84
	2048	15.58	16.20	15.30	102518	36.63	55089.06	32.35
torso1	32	2.36	2.36	2.35	6	2.64	5.33	2.58
	64	2.38	2.39	2.37	11	2.94	6.30	2.86
	128	2.41	2.44	2.40	26	3.63	19.87	3.31
	256	2.49	2.55	2.45	128	3.84	45.34	4.62
	512	2.62	2.75	2.55	427	7.10	249.26	6.95
	1024	2.88	3.18	2.76	1862	11.84	657.00	11.19
	2048	3.40	4.02	3.16	7485	18.86	1329.05	19.65
Averages normalized w.r.t. RB1 times								
	32	1.00	1.01	1.00	3	1.26	2.58	1.18
	64	1.00	1.02	0.99	6	1.47	7.70	1.37
	128	1.00	1.03	0.99	46	1.75	24.12	1.69
	256	1.00	1.06	0.97	269	2.61	198.60	2.29
	512	1.00	1.09	0.95	796	3.18	548.24	3.07
	1024	1.00	1.14	0.93	1953	4.09	1160.17	4.20
	2048	1.00	1.21	0.91	4074	5.83	1595.60	5.03

expected. Although the best imbalance values decrease with increasing R, the decreases are quite small, especially for large P. Moreover, for a fixed R, the relative difference between the best and average imbalance values decreases with increasing P.

Table 7
Geometric averages of percent load imbalance ratios for R randomly ordered processor chains for the volume rendering dataset

	$R = 10$		$R = 100$		$R = 1000$		$R = 10000$	
P	best	avg	best	avg	best	avg	best	avg
32	0.043	0.051	0.039	0.051	0.034	0.052	0.033	0.052
64	0.091	0.104	0.085	0.126	0.081	0.108	0.074	0.118
128	0.201	0.215	0.190	0.214	0.176	0.213	0.170	0.213
256	0.395	0.425	0.378	0.425	0.377	0.425	0.363	0.425
512	0.822	0.849	0.803	0.852	0.802	0.853	0.772	0.853
1024	1.638	1.686	1.599	1.680	1.593	1.680	1.578	1.681
2048	3.369	3.449	3.335	3.447	3.290	3.446	3.250	3.445

Table 8
Geometric averages of percent load imbalance ratios for R randomly ordered processor chains for the sparse matrix dataset

	$R = 10$		$R = 100$		$R = 1000$		$R = 10000$	
P	best	avg	best	avg	best	avg	best	avg
32	0.219	0.467	0.135	0.405	0.061	0.412	0.053	0.425
64	0.436	0.673	0.379	0.827	0.305	0.901	0.231	0.910
128	1.110	2.164	0.960	2.154	0.885	2.278	0.815	2.250
256	8.743	9.148	8.534	9.191	8.377	9.221	8.317	9.858
512	29.024	30.386	28.593	30.266	28.182	30.186	27.883	30.175
1024	66.090	67.322	65.464	67.427	64.720	67.370	64.176	67.370
2048	139.262	140.655	137.273	140.715	136.582	140.684	135.486	140.699

These experimental findings show that processor ordering is important for solution quality only for small P. This is expected since the variance among processor speeds is low, unlike the variance among task weights. These experimental results also show that exact CCP algorithms can serve as an effective heuristic for the CP problem.

7 Conclusions

We studied the problem of one-dimensional partitioning of nonuniform workload arrays with optimal load balancing for heterogeneous systems. We investigated two cases: chain-on-chain partitioning, where a chain of tasks is partitioned onto a chain of processors; and chain partitioning, where the task chain is partitioned onto a set of processors (i.e., permutation of the processors is allowed). We showed that chain-on-chain partitioning algorithms

for homogenous systems can be revised to solve this partitioning problem for heterogeneous systems, without altering computational complexities of these algorithms. We proved that the chain partitioning problem is NP-complete, and empirically showed that exact CCP algorithms can serve as an effective heuristic, for the CP problem. Our experiments proved the effectiveness of our techniques, as the exact algorithms work much better than heuristics, and balanced work decompositions can be achieved even for high numbers of processors.

References

[1] NASA advanced supercomputing division (NAS) dataset archive, http://www.nas.nasa.gov/Research/Datasets/datasets.html.

[2] B. B. Cambazoglu, C. Aykanat, Hypergraph-partitioning-based remapping models for image-space-parallel direct volume rendering of unstructured grids, IEEE Transactions on Parallel and Distributed Systems 18 (1) (2007) 3–16.

[3] H.-A. Choi, B. Narahari, Algorithms for mapping and partitioning chain structured parallel computations, in: International Conference on Parallel Processing, 1991.

[4] T. Davis, University of Florida Sparse Matrix Collection, http://www.cise.ufl.edu/research/sparse/matrices, NA Digest, vol. 97, no. 23 (June 1997).

[5] K. D. Devine, B. Hendrickson, E. G. Boman, M. M. S. John, C. Vaughan, Zoltan: a dynamic load-balancing library for parallel applications – user's guide, Tech. Rep. SAND99-1377, Sandia National Libraries (1999).

[6] M. R. Garey, D. S. Johnson, Computers and Intractability; A Guide to the Theory of NP-Completeness, W. H. Freeman & Co., New York, NY, USA, 1990.

[7] M. P. Garrity, Raytracing irregular volume data, in: VVS '90: Proceedings of the 1990 workshop on Volume visualization, ACM Press, New York, NY, USA, 1990.

[8] H. Kutluca, T. M. Kurç, C. Aykanat, Image-space decomposition algorithms for sort-first parallel volume rendering of unstructured grids, The Journal of Supercomputing 15 (1) (2000) 51–93.

[9] V. J. Leung, E. M. Arkin, M. A. Bender, D. Bunde, J. Johnston, A. Lal, J. S. B. Mitchell, C. Phillips, S. S. Seiden, Processor allocation on Cplant: Achieving general processor locality using one-dimensional allocation strategies, in: CLUSTER '02: Proceedings of the IEEE International Conference on Cluster Computing, IEEE Computer Society, Washington, DC, USA, 2002.

[10] F. Manne, B. Olstad, Efficient partitioning of sequences, IEEE Transactions on Computers 44 (11) (1995) 1322–1326.

[11] S. Miguet, J.-M. Pierson, Heuristics for 1D rectilinear partitioning as a low cost and high quality answer to dynamic load balancing, in: HPCN Europe '97: Proceedings of the International Conference and Exhibition on High-Performance Computing and Networking, Springer-Verlag, London, UK, 1997.

[12] D. M. Nicol, Rectilinear partitioning of irregular data parallel computations, Journal of Parallel and Distributed Computing 23 (2) (1994) 119–134.

[13] J. R. Pilkington, S. B. Baden, Dynamic partitioning of non-uniform structured workloads with spacefilling curves, IEEE Transactions on Parallel and Distributed Systems 7 (3) (1996) 288–300.

[14] A. Pınar, C. Aykanat, Sparse matrix decomposition with optimal load balancing, in: HIPC '97: Proceedings of the Fourth International Conference on High-Performance Computing, IEEE Computer Society, Washington, DC, USA, 1997.

[15] A. Pinar, C. Aykanat, Fast optimal load balancing algorithms for 1D partitioning, Journal of Parallel and Distributed Computing 64 (8) (2004) 974–996.

[16] P. Shirley, A. Tuchman, A polygonal approximation to direct scalar volume rendering, in: VVS '90: Proceedings of the 1990 workshop on Volume visualization, ACM Press, New York, NY, USA, 1990.